# A Tool for Circumventing Modern Techniques of Authorship Attribution

Research proposal

Alden Page

December 2, 2014

**Abstract**

Linguists have developed incredibly effective methods of determining the authorship of anonymous documents with computer assistance. While this has helped answer a number of intriguing questions about disputed authorship, it can also be used as a threat to privacy and anonymity. As a result, there is a surge of interest in a field that researchers have deemed "adversarial stylometry," in which computers are used to assist in the obfuscation of writing style. As of the time of the writing of this document, few tools exist for effectively obfuscating writing style, and none are especially friendly for non-technical users. In order to rectify this, I propose to develop an open source adversarial stylometry application that addresses the failings of already existing tools in this domain.

## 1 Introduction

Many people have an interest in preserving their anonymity out of fears of censorship, political persecution, and personal safety[1][2][3]. While a wide variety of tools exist for preserving network level anonymity (i.e., hiding IP addresses and encrypting data), there are few avenues for hiding stylistic information. Given the increasing number of sophisticated authorship attribution techniques employable by large corporate or state entities, avoiding authorship attribution should be a major concern of anybody wishing to preserve their anonymity.

### Introduction to stylometry

Stylometry is the practice of extracting *linguistic features* of style from a body of written work called a *corpus.* Stylometry is of interest due to its applicability in the field of authorship attribution, in which researchers attempt to determine the original author of a document for which authorship is disputed[4]. While the term "stylometry" encompasses a wide variety of stylistic

traits not necessarily found in writing, for purposes of this proposal, stylometry and authorship attribution will be used interchangeably.

In the context of stylometry, the term "linguistic corpus" tends to refer to a collection of documents by a number of different authors, and "linguistic features" refers to distinct reoccurring patterns in an author's corpora. The number of possible linguistic features in a document is quite large, but only a small number of them have been proven to be useful for authorship attribution in an empirical manner. The complexity of linguistic features ranges anywhere from simple word length metrics to more computationally expensive "deep" metrics, such as syntactic phrase structure. In either case, linguists and computer scientists have made extensive use of machine learning in order to successfully extract, classify, and rank the importance of features for purposes of attribution.

In response to stylometry, there is an emerging field of research focusing on evading modern methods of authorship attribution, deemed *adversarial stylometry* by Brennan et al[5]. The concept of adversarial stylometry proposes the idea that an attacker aware of stylometric features could modify his or her writing style in order to partially or entirely invalidate the use of authorship attribution analysis through either *imitation*, *translation*, or *obfuscation*, with machine-assisted obfuscation serving as the current most promising method of circumventing stylometry.

## Importance and potential impact

By combining network-level privacy tools like Tor[1] with an effective tool for obfuscating stylistic features, one could have at least a reasonable assurance of total anonymity, both from a behavioral and network perspective.

For most security applications, obfuscating textual style may strike some as borderline paranoid. Indeed, making use of adversarial stylometry is likely overkill for most security applications, and patently absurd in other domains. Whereas other existing privacy-enhancing technologies like encryption have their place in the daily lives of millions of internet users, adversarial stylometry should be reserved for situations in which being identified is a dire risk to personal safety or way of life.

## Motivating analogy

While situations warranting stylometric obfuscation are few, they are often of national political significance or involve protection of the principles of free speech. Consider Bob, a theoretical citizen of a repressive regime. Bob has distributed anonymous essays that express dissatisfaction with the state establishment. In addition, Bob has been previously imprisoned for complaining about working conditions. Walter, a state actor whose job is to identify and intern political dissidents, has obtained a copy of the essay and sets about determining the authorship of the document. Walter performs stylometric analysis on the document and, using a database of corpora of previous political agitators, finds that approximately five former political offenders tend to use the

[1]Tor is a software suite designed to assist in hiding the user's location and activity while using the internet. A recent leak revealed that the NSA described Tor as "The king of high-secure, low-latency anonymity."

particular configuration of function words and n-grams found in the anonymous document. By augmenting this information with the use of traditional investigation techniques such as interviews, interrogation, and intimidation, Walter correctly deduces that Bob wrote the incriminating document and turns him over to the police. Perhaps Walter does not even bother with this step, and simply imprisons the five prime suspects. Had Bob made use of adversarial stylometry tools, he likely would have avoided apprehension by the authorities.

The previous analogy makes some assumptions that likely initially strike the reader as unrealistic. How could Walter have access to Bob's previous linguistic corpus? With the proliferation and public availability of social media, it is less difficult than one might think to assemble a corpus for a particular individual. Bob must have had a public blog completed unrelated to politics, such as a do-it-yourself or cooking themed blog. Since Bob has a record of political agitation, his online activity is archived by the state, and his blog is added to Walter's available corpora. Application of modern techniques of stylometry could potentially associate Bob with the anonymous document, particularly in light of authorship attribution becoming increasingly effective with large corpora and smaller bodies of text.

More mundane examples of the applicability of adversarial stylometry are not difficult to conjure up. Perhaps Alice has witnessed corruption in her company and wishes to write an anonymous account to the local newspaper, hoping that her boss Mallory does not use her work emails as a linguistic corpora to expose her identity, endangering her employment status and likely her livelihood.

# 2 Background and Related Work

## Early work

Pioneering work in stylometry focused on authors of historical significance, with a notable example being studies on the stylistics of the Federalist Papers[6]. 12 of the 85 articles of the Federalist Papers were of disputed authorship, though it was always suspected that either Alexander Hamilton or James Madison wrote them. By taking note of small differences in the choice of function words and making use of Bayesian inference, Mosteller and Wallace were able to say with a high degree of certainty that James Madison was the author of all 12 of the disputed essays. Later studies using more sophisticated machine learning models have overwhelmingly agreed with these findings. Another frequent application of stylometry is the authorship of Shakespeare's work, though there is currently little scholarly consensus in this particular domain.

There are countless quantifiable stylistic features that linguists have made use of in studies on stylometry, with much research focusing on the occurrence of *n-grams*[2] and *function words*[3]. Only a small number of these features are known to be usable in authorship attribution. A tool hoping to defeat stylometry must therefore focus on these particular features. Stamatos provided a survey

[2]n-grams are, informally, groups of characters of $n$ length. Note that n-grams include spaces, often times represented as underscores. For instance, the first 3-gram (or trigram) from this sentence is given by "_Fo".

[3]Function words are connecting words in sentences with little or no lexical information, such as the words "the", "a", and "and".

of effective methods for authorship attribution, providing context and focus for an adversarial stylometry tool[4].

## Current state of research in authorship attribution

Studies involving stylometry typically involve a sample size of 10 to 300 authors. In light of constantly improving computational capacity and the public availability of large corpora on the internet, a few studies have attempted to use stylometry on corpora with thousands of potential authors. Narayayan et al. recently showed that internet-scale authorship attribution is feasible to a surprising degree[7]. This study made use of proven stylometric methods on a sample size of 100,000 potential authors. The classifier succeeded in identifying authors approximately 20% of the time. Prolific writers with more than 40 posts on their blog could be identified as much as 35% of the time. While this amount of accuracy is too low to identify authors singlehandly, the classifier can be used in combination with manual analysis of documents to great effect in identifying authors, since the pool of possible authors has been shrunk from 100,000 to 100-200 by a ranking algorithm. Furthermore, a few adjustments to the classifier could improve accuracy to up to 80% by adjusting the sample size slightly. While 20 to 35% accuracy is not a dependable number on its own, large-scale stylometric analysis could be used in combination with other security leaks in order to unmask an anonymous person's identity. This shows that it is quite possible to apply traditional means of stylometry on enormous sample sizes, and further illustrates the importance of proper tools for adversarial stylometry in the preservation of anonymity.

Further challenging assumptions about the requisite number of authors and available works to perform authorship attribution is a study showing that stylometric analysis can be effectively used on Twitter[4] posts[8]. Bhargava et al. showed that it is possible to attain 95% accuracy in authorship attribution with approximately 200 posts in a pool of 10 to 30 users. Given that many active Twitter users have thousands of posts, it appears that stylometry can be performed quite effectively on social media platforms, even with exceptionally short and unrelated samples.

It has been shown that simply imitating another author's writing style significantly impairs the use of some of the more superficial features used in authorship attribution, such as average word length or vocabulary richness[9]. Imitation also works well on more sophisticated forms of stylometric analysis, although to a lesser degree. The problem with the imitation approach is that it is difficult to apply to writing samples of extended length without ever allowing one's own writing style to "leak" through, possibly resulting in the inadvertent identification of the original author.

Machine translation from one language to another back to the parent language is not an effective method of hiding writing style[9]. It comes at great cost to the readability of the writing and still fails to hide linguistic features to any reasonable degree. It seems that linguistic features persist through any number of rounds of translation.

[4]Twitter is a social media platform in which users submit posts ("tweets") consisting of no more than 140 characters.

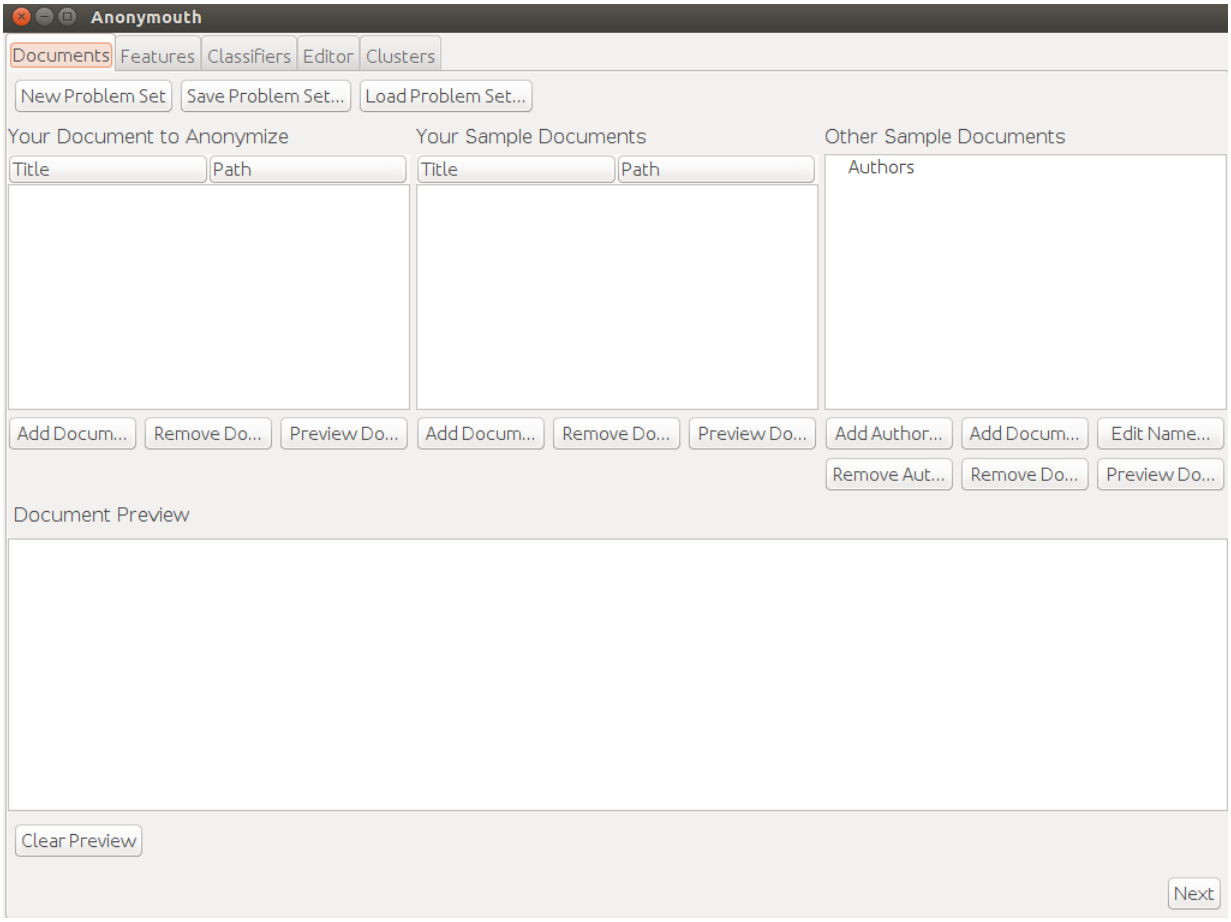


Figure 1: A screenshot of Anonymouth's main screen.

## Current state of adversarial stylometry

The only adversarial stylometry tool that is available to the public is an application called Anonymouth, a tool developed at Drexel University designed primarily with linguistic researchers in mind[5]. Previous research also suggests that the feature set that Anonymouth focuses on may be flawed[10]. While Anonymouth and its developers deserve praise for their pioneering work in adversarial stylometry, Anonymouth is primarily a proof of concept designed for research use rather than a practical tool for obfuscating users' identities. There are a number of self-admitted usability issues, from the instability of the application to expecting users to hunt down and input work by other authors into the application.

One could characterize Anonymouth as "fragile." The instability of Anonymouth is quite serious, particularly on Linux distributions. Over the course of but a few simple preliminary experiments using corpora that was bundled with the software, Anonymouth crashed with little encouragement, and often threw warnings to the user. While these errors may have been navigable for researchers, end users should not have to deal with constant crashing. Of course, Anonymouth is currently in version 0.0.1, but there have not been any updates for the past year and a half, suggesting that the software is no longer in development.

Additionally, Anonymouth does not attempt to automatically obfuscate features when possible.

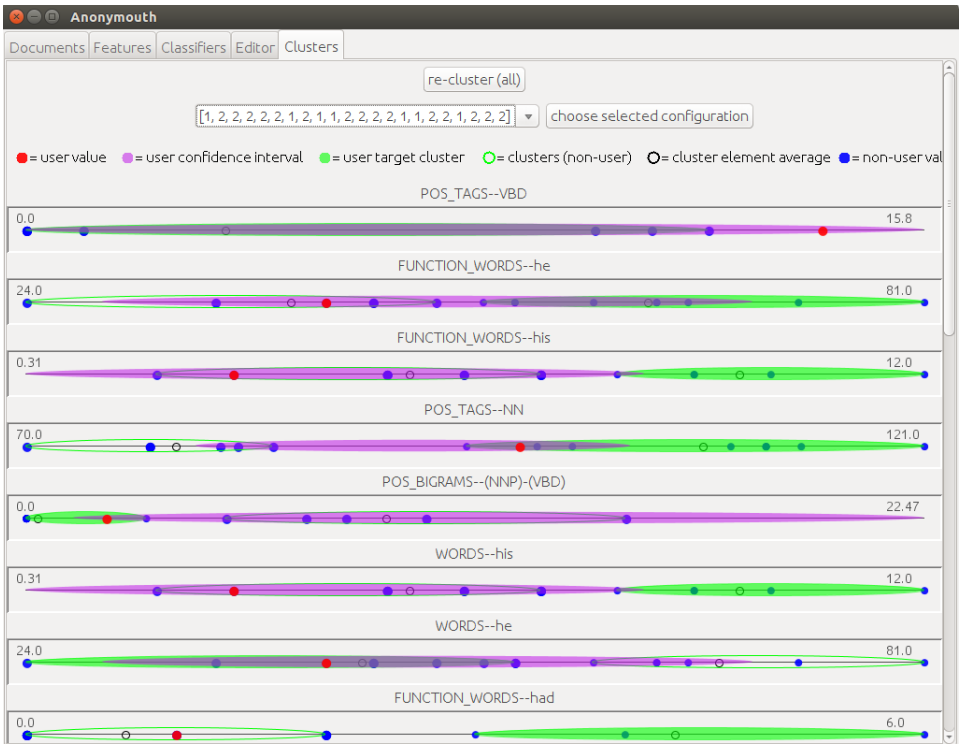


Figure 2: The particularly baffling clustering screen in Anonymouth.

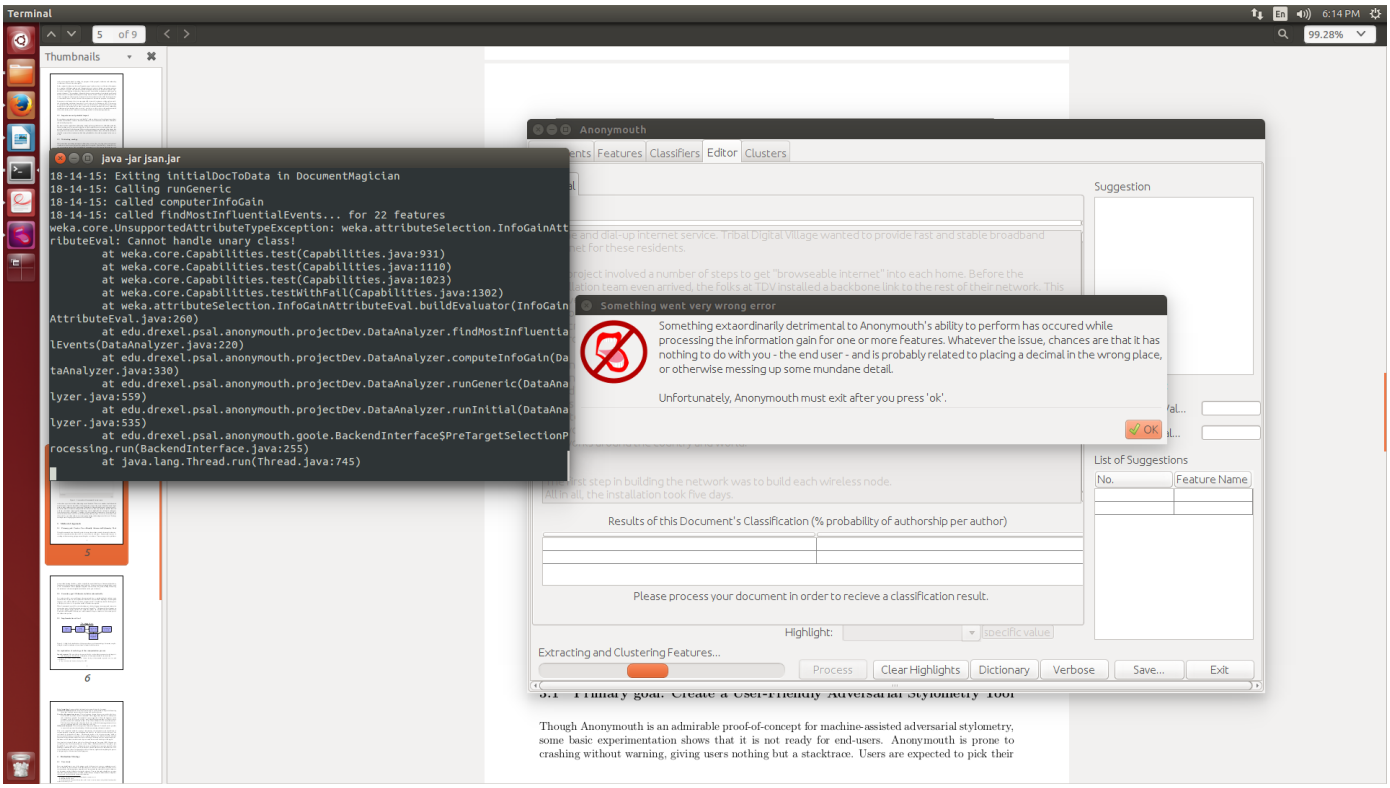


Figure 3: A screenshot illustrating Anonymouth's instability.

For instance, it may be possible to "jumble" function words in a document automatically, eliminating a class of highly effective stylometric techniques from the adversary's toolbox[11]. A number of other stylometric features may be vulnerable to automatic obfuscation as well.

It is worth exploring which portions of text can be automatically obfuscated by a machine without user intervention during the development of this tool. Perhaps existing natural language generation tools will be helpful in this regard.

# 3 Method of Approach

## Primary goal: Create a User-Friendly Adversarial Stylometry Tool

Though Anonymouth is an admirable proof-of-concept for machine-assisted adversarial stylometry, some basic experimentation shows that it is not ready for end-users. Anonymouth is prone to crashing without warning, giving users nothing but a stacktrace. Users are expected to pick their own machine learning classifier, a task not suitable for non-technical users. Anonymouth produces long lists of stylistic features with little visual feedback. Certain visual features simply do not work at all – Anonymouth fails to highlight dangerous reoccurrences of n-grams, making obfuscating the stylistics of a document significantly harder on the part of the user.

Often times, it seems that many of Anonymouth's biggest flaws are easily fixed with more thorough documentation, reduction of unnecessary choices for the user, or plain intuition. The proposed tool will make the user the first priority, at the cost of usefulness as a platform for research. For instance, the tool only needs to support a single classifier, whereas Anonymouth supports about a dozen for purposes of research. Anonymouth also requires the user to input documents from many different authors, while the proposed tool could simply contain a pre-loaded database of corpora from other authors in order to save the user time.

## Secondary goal: Obfuscate stylistics automatically

In a perfect world, a user could input a document and have a machine hide the stylistic traits automatically. In practice, this requires the machine to have extensive knowledge of the English language, and could be difficult to achieve programatically.[5] Nonetheless, should all other goals of this project be met, it is a problem worthy of further investigation.

When Anonymouth was still in active development, the development team expressed interest in automating parts of the obfuscation process using ConceptNet.[6] Unfortunately, development on the project stopped before this idea was taken any further. In combination with the Stanford Parser and the SimpleNLG library, one could reasonably hope to automate at least some part of the obfuscation process.

[5]This is a concern that Andrew McDonald, a former developer of Anonymouth, expressed to me in e-mail correspondence.

[6]A large "common sense" database developed by MIT.

## Implementation of tool

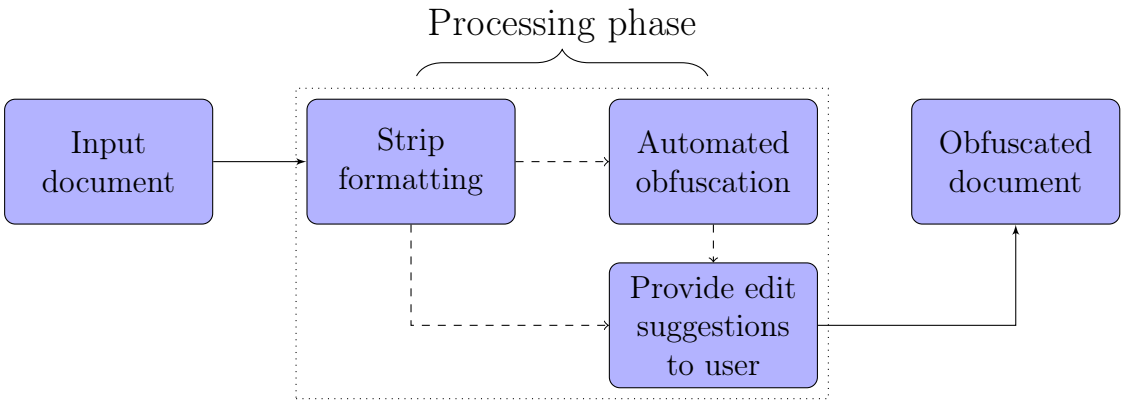


Figure 4: A high-level visualization of the proposed process of anonymizing a document, emphasizing the tentative inclusion of some form of automated obfuscation.

**Input document**
The user selects a document that he or she wishes to anonymize and inputs it into the program. For the sake of simplicity and flexability, all documents should be in plain text.

**Strip formatting**
Any non-stylistic information is removed from the document.

**Automated obfuscation**
The tool makes some changes to the document in hopes of obfuscating the author's stylistic features, thereby leaving less work for the user.

**Provide edit suggestions to user**
The tool suggests changes the user can make to the document in order to promote stylistic anonymity. These suggestions should be as transparent and actionable as possible. An example of a poor suggestion is "lower the frequency of the occurrence of the 'de' bigram," where a more helpful suggestion might say "replace these particular words with synonyms in order to lower the occurrences of 'de' in this document." The latter's specificity provides direction for the user, while the former suggestion may leave the user wondering what the tool is asking the user to do.

**Obfuscated document**
The tool outputs the obfuscated document. It should not be possible to accurately ascertain the authorship of the document using stylometric analysis.

Figure 5: An explanation of each stage of the anonymization process described in the previous chart.

This tool is essentially intended to replicate functionality of Anonymouth, with the addition of features intended to improve user-friendliness and stability.

In order to faciliate the goal of simplicity and usability, the tool should be developed in Python. This domain involves a lot of text processing, which is quite natural in Python as compared to most compiled languages. In addition, Python's minimal syntax and high readability makes code

easier to manage, a desirable trait in a one-person project. Judicious use of Python's modular and object-oriented features will also help in this regard.

One drawback of using Python is that few Natural Language Processing (NLP) libraries are available natively. Most relevant libraries, such as Weka[7], the Stanford Parser[8], ConceptNet, and SimpleNLG[9], are written in Java. Fortunately, many of the libraries in question provide Python bindings. Additionally, the `javabridge` interface allows Python users to instantiate the Java Virtual Machine and interact with unsupported Java libraries, significantly simplifying the process of integrating Java libraries into Python programs.

# 4 Evaluation Strategy

## User study

Since user-friendliness is one of the primary goals of this project, a survey confirming ease-of-use is essential in the evaluation of the effectiveness of the tool. Additionally, user feedback could be priceless in future work with this project.

Users will be asked to go through the process of obfuscating the stylistics of a (preferably user-supplied) document using the proposed tool. No aid besides included project documentation and/or a first-time user wizard will be provided. After the user has completed his or her attempt to obfuscate a document, the user will be asked a series of questions, including, but not limited to:

- Did you successfully anonymize your document? (y/n)
- On a scale of 1 to 10, with 1 being the least difficult and 10 being the most difficult, how hard was the process of using this tool?
- If you encountered difficulties using the tool, what was the biggest issue?
- Did the program ever crash? (y/n)
- Did the program ever freeze for lengthy periods of time? (y/n)
- On a scale of 1 to 10, with 1 being extremely responsive and 10 being highly unresponsive, how responsive was the application to your inputs? (For example, did you find the application froze when you clicked certain buttons, or caused your mouse pointer to get stuck? If you answered yes to either, then the program was unresponsive to some degree.)
- Do you see any areas of improvement for this tool?

Users in this study should have no prior experience with stylometric tools, yet still be technically apt enough to successfully use a new piece of software on short notice. Students who have successfully completed an introductory computer science course would be prime candidates.

[7] A Machine Learning library that Anonymouth makes extensive use of.

[8] An English language parser

[9] A Natural Language Generation library that could be used to rewrite sentences, hopefully obfuscating their stylistic traits in the process.

## Stylometric obfuscation effectiveness study

In order to be useful, this tool must obfuscate stylistic traits effectively. In order to verify this, one should ensure that the output document from the tool is actually properly obfuscated. This can be confirmed with existing tools like JStylo[10]. Through this, we seek to answer two questions:

1. Can an expert user of this tool successfully anonymize a document?
2. Can a first-time user successfully anonymize a document?

It is simple enough to answer both of these questions using the same criteria as the Anonymouth project did. The following stylistic criteria are taken account of in the 9-feature set described by Brennan et al. in the Anonymouth project[5]:

**Unique Words Count**
The Number of unique words in the document after removing punctuation and unifying case.

**Complexity**
Ratio of unique words to total number of words in the document.

**Sentence Count**
The number of sentences in the document.

**Average Sentence Length**
The total number of words divided by the total number of sentences.

**Average Syllables in Word**
Number of syllables per word divided by the number of words.

**Gunning-Fog Readability Index**
The Gunning-Fog readability index is given by:

$$0.4\left[\left(\frac{totalwords}{totalsentences}\right)+100\left(\frac{totalcomplexwords}{totalwords}\right)\right],$$

where complex words are words with 3 or more syllables.

**Character Space**
The total number of characters in the document, including spaces.

**Letter Space**
The total number of letters (excluding spaces and punctuation.)

**Flesch Reading Ease Score**
The Flesch reading ease score is given by:

$$206.835-1.015\left(\frac{totalwords}{totalsentences}\right)-84.6\left(\frac{totalsyllables}{totalwords}\right).$$

An effective adversarial stylometry tool should, at the very least, considerably alter the occurrence of these features such that it is difficult to perform any kind of useful authorship analysis.

More specifically, consider the use of the JStylo authorship attribution framework. Assume JStylo is able to guess the authorship of an input document with 90% certainty under the Greenstadt 9-feature set. After the tool is used on this same document, the authorship certainty should be considerably reduced. 50% certainty after obfuscation should be a reasonable goal.

[10] A tool developed by the Anonymouth team for purposes of applying stylometric analysis on text.

# 5 Research Schedule

| Task | Begin Date | End Date | Notes |
|---|---|---|---|
| Implement tool | January 10 | March 10 | Implement adversarial stylometry tool, with first priority being getting the application in at least a minimally usable state. |
| Complete at least two chapters | Feb 2 | Feb 28 | Finish first two chapters of proposal. |
| Perform user and effectiveness studies | March 11 | March 13 | Collect a reasonable number of usability surveys. Evaluate output documents to ensure that user successfully obfuscated their document. I expect that all data could be collected in a single day, and analysis could be completed the following day, but have allocated an extra day just in case something goes awry. |
| Refine draft | March 13 | April 1 | Continue working on remaining chapters in senior thesis. |
| Submit unbound thesis | April 2 | – | – |
| Oral defense | As early as is feasible | No later than April 21 | In order to maximize time for making suggested changes to my thesis, I would like to complete my oral defense at the earliest possible date. The date that I register for my defense is conditional on workload and the dates of exams in the month of April. |

# 6 Conclusion

For the sake of the preservation of anonymity, it is essential that a high quality and easy-to-use tool for obfuscating writing style becomes widely available. As of 2014, the only available

tool that provides adversarial stylometry functionality is Anonymouth, which leaves much to be desired in the realm of usability and stability. In order to rectify this situation, I will develop a lightweight tool in Python that aims to address the weaknesses of Anonymouth. In order to verify the effectiveness of the tool, two studies will be conducted: one ensuring the application is user-friendly, and one ensuring that the stylistic obfuscation process is effective.

# References


[1] The Anonymous SEC Whistleblower Award of $14M Is A Game Changer. URL http://www.forbes.com/sites/walterpavlo/2013/10/03/the-anonymous-sec-whistleblower-award-of-14m-is-a-game-changer/, 2013.

[2] How a computer program helped reveal j.k. rowling as author of a cuckoo's calling. URL http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/, 2013.

[3] Carole E. Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *IJDE*, 2005.

[4] Efstathios Stamatatos. Survey of modern authorship attribution methods. In *Journal of the American Society for Information Society and Technology*, pages 539–556. American Society for Information Science and Technology, 2008.

[5] AndrewW.E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. Use fewer instances of the letter "i": Toward writing style anonymization. In Simone Fischer-Hubner and Matthew Wright, editors, *Privacy Enhancing Technologies*, volume 7384 of *Lecture Notes in Computer Science*, pages 299–318. Springer Berlin Heidelberg, 2012.

[6] G. S. Watson. Inference and disputed authorship: The federalist. *The Annals of Mathematical Statistics*, 37(1):pp. 308–312, 1966.

[7] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Eui Chul, Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *In In Proceedings of the 33rd conference on IEEE Sympsoium on Security and Privacy. IEEE*, 2012.

[8] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric analysis for authorship attribution on twitter. In Vasudha Bhatnagar and Srinath Srinivasa, editors, *Big Data Analytics*, volume 8302 of *Lecture Notes in Computer Science*, pages 37–47. Springer International Publishing, 2013.

[9] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22, November 2012.

[10] Carole E. Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8, 2001.

[11] Shlomo Argamon and Shlomo Levitan. Measuring the usefulness of function words for authorship attribution. In *In Proceedings of the 2005 ACH/ALLC Conference*, 2005.